Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Academic Year (2022-23)
Year: 3 Semester: V

Program: Minor in Data Science **Max. Marks: 75**
Subject: Foundation of Data Analysis **Time: 10: 30 am to 1:30 pm**
Date: 3/01/2023 **Duration: 3 Hours**

REGULAR EXAMINATION

Instructions: Candidates should read carefully the instructions printed on the question paper and on the cover page of the Answer Book, which is provided for their use.
(1) This question paper contains **03** pages.
(2) All Questions are Compulsory.
(3) All questions carry equal marks.
(4) Answer to each new question is to be started on a fresh page.
(5) Figures in the brackets on the right indicate full marks.
(6) Assume suitable data wherever required, but justify it.
(7) Draw the neat labelled diagrams, wherever necessary.

Question No.		Max. Marks
Q1 (a)	i. Calculate the arithmetic mean for the following weight recorded to the nearest grams of 60 apples picked out at random from a consignment are given below:	[05]

Weights (grams)	Frequency
65-84	09
85-104	10
105-124	17
125-144	10
145-164	05
165-184	04
185-204	05

OR

	ii. Answer the following: a. Calculate mode for ungrouped data. X_i = 2 3 8 4 6 3 2 5 3 b. Calculate the median for the following grade points obtained by 10 practitioners are given. X_i = 45 32 37 46 39 36 41 48 36 50	[03] [02]
Q1 (b)	i. Explain the relative measures of dispersion in statistics. ii. Describe the steps used to calculate the score for a given percentile P. iii. The scores of students are 3, 5, 7, 8, 9, 11, 13, 15. What is the score of the 25th percentile?	[05] [02] [03]
Q2 (a)	1. Answer the following: i. Explain in detail the importance of data cube. ii. Sketch a 3D data cube for sales of an organization. iii. Sketch a Star schema of sales where Dealer, Model, Date, Product, Branch are dimension table & revenue is a fact table with Units sold and total revenue. Assume necessary attributes.	[03] [03] [04]
	OR	

	2. Answer the following: i. Explain various OLAP models in data warehouse. ii. Consider the following data illustrating temperature of certain weeks: <table><tr><th>Temperature</th><th>cool</th><th>mild</th><th>hot</th></tr><tr><td>Week1</td><td>2</td><td>1</td><td>1</td></tr><tr><td>Week2</td><td>2</td><td>1</td><td>1</td></tr></table> Suppose we want to set daily temperature from the above data. Specify the operation used to create a new dataset.	[05] [05]
Q2.(b)	Select the appropriate answers for the below questions and justify your answer. i. Data visualization tools provide an accessible way to see and understand ._____ in data. a) trends b) outliers c) patterns d) All of the above ii. The charts that are helpful in making comparisons with features are _____. a) Bar charts b) Column charts c) Pie charts d) Both bar and column charts iii. A data visualization that updates in real-time and gives multiple outputs is called as _____. a) Dashboard b) Metrics table c) Data table d) None of the above iv. For creating variable size bins we use _____. a) Sets b) Groups c) Calculated Fields d) Table Calculations v. Which graphs are generally used to show the re4lationship among the variables? a) Bar graph b) Line graph c) Scatter Plot d) Maps	[01] [01] [01] [01]
Q3 (a)	i. What is the need of data preprocessing? Explain the various techniques of data cleaning with a suitable example. **OR** ii. Given the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. a. Use min-max normalization to transform the value 35 for age onto the range [0.0,1.0]. b. Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years. c. Use normalization by decimal scaling to transform the value 35 for age.	[05] [05]

Q3 (b)

i. Explain the relation between arithmetic, geometric and harmonic mean with suitable example. [10]

OR

ii. Explain the various absolute measure of dispersion. [05]

iii. Find the range and coefficient of range of the following: 25, 67, 48, 53, 18, 39, 44. [05]

Q4 (a)

1. Answer the following: [03]

Calculate the Pearson correlation coefficient for the Marks obtained by 5 students in algebra and trigonometry as given below:

Science	16	15	12	10	8
Geometry	11	18	10	20	17

ii. Find the SNR for the following data: 6, 24, 6, 14, 10. [02]

iii. What would be the standard deviation of the data if the mean is 45 and SNR is 39.5? [02]

OR

2. Compute the Spearman's rank correlation coefficient of the following data. Also specify if the ρ value is perfect association of rank or no. [07]

Rank in History (X)	3	5	1	6	7	2	8	9	4
Rank in Geography (Y)	5	3	2	6	8	1	7	9	4

Q4 (b)

i. Given the following data (in increasing order) for the attribute age: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34. Calculate the following using a bin size of 3:

a) Use smoothing by bin mean to smooth these data. [08]

b) Use smoothing by bin boundary to smooth these data.

Q5 (a)

Solve any two.

i. Explain with an example the difference between correlation and causation. [05]

ii. Explain various Multivariate methods in feature engineering. [05]

iii. If two variables have a high correlation with a third variable, does this convey they will also be highly correlated? Is it even possible that A and B are positively correlated to another variable C? Is it possible that A and B are negatively correlated with each other? [05]

iv. Explain extreme values and limits in examining the relationship in correlation. [05]

Q5 (b)

i. Calculate the correlation coefficient for the following height (in inches) of mother (X) and their daughter (Y) using arbitrary origin method of Karl Pearson's coefficient. [05]

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71